1er avance en el nivel económico

***1er objetivo*:** *Entender los formatos de la base de datos del Banco Mundial utilizando como referencia los registros sobre Hong Kong.*

Base de datos

En el momento que se presenta este avance se especifica que únicamente se han trabajado datos provenientes de los archivos en formato API CSV del Banco Mundial en español.

- La base de datos del país tiene 1438 indicadores económicos, socio-económicos, etc. (La mayoría en español)
- Cada indicador es una serie de datos a través del tiempo.
- Las series de tiempo que se pretenden registrar son desde 1960 hasta 2019.
- El tipo de registro, hasta el momento, es completamente numérico. (Una variable continua)
- Los registros están medidos según las dimensiones del indicador: Enteros con decimales, pueden ser porcentajes o no. (Lo especifica el indicador)
- Series enteras: series de datos completamente registradas en los años que se especifican.
- Series nulas: series de datos completamente nulas (sin registro en ningún año)
- Series mixtas: series con al menos un valor registrado (*no nulo*).

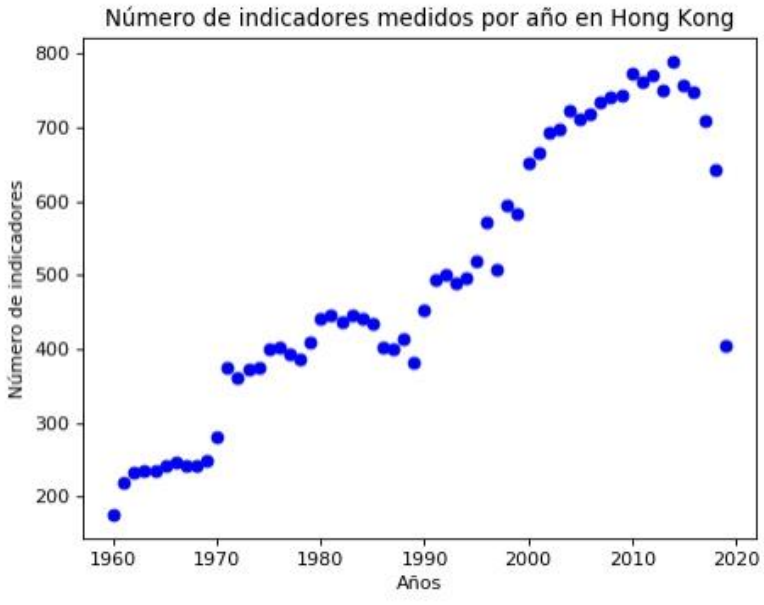
Las series mixtas existen de tres maneras:

1) Intermitente: series de datos donde no se registraron determinados años en medio de registros existentes.
2) Posterior: series registradas después de determinado año. Antes de él no existían registros de determinado indicador.
3) Mixta: series con los dos casos anteriores.

La siguiente gráfica ilustra que cada año tiene una medición distinta de indicadores. Los indicadores considerados para la suma tienen al menos un registro no vacío (*no nulo*), es decir, son al menos mixtas.

Esta gráfica es parte de la primera conclusión al familiarizarse con la base de datos.

```
El máximo de indicadores registrados es 790
El mínimo de indicadores registrados es 175
```



¿Qué nos dice? (Observaciones)

1) El mínimo de indicadores (series de datos) con al menos un valor no nulo se encuentra al inicio de la serie de tiempo. En 1960. Es de apenas 175 indicadores.
2) El máximo de indicadores (series de datos) con al menos un valor no nulo se encuentra en los últimos años. Es de 790.
3) La tendencia del registro suele ser creciente.
4) Los últimos años empiezan a decrecer en el número de indicadores medidos.
5) Hay tres puntos de inflexión en la gráfica.

Conclusiones importantes

Como el número de indicadores totales a registrar es 1438, podemos notar que hay un gran número de series nulas por año dado el bajo registro. Al buscar un indicador de nuestro interés, hay alrededor de un 40% de probabilidad de que sea una serie nula.

2do objetivo: *Encontrar los indicadores más relevantes usando las palabras clave dadas en la rúbrica del nivel económico y crear un marco de datos con ellos*.

Creando un marco de datos limpio y útil.

Al saber de la existencia de series nulas de datos, tendríamos dos opciones para manejar la base.

1) Borrar todas las series nulas desde el comienzo.
2) Dejar las series nulas y limpiar conforme busquemos indicadores.

Se eligió la opción dos con el propósito de tener idea de todos los indicadores que pudiésemos encontrar. En dado caso de que tengamos interés por un indicador, pero la serie de datos sea nula, podemos recurrir a otra forma de encontrar ese valor. De haber elegido la opción uno, estaríamos cerca de 40% limitados en cuanto al conocimiento de los posibles indicadores.

- De los 72 indicadores de los cuales se mostró interés en un inicio, 52 (únicos) de ellos resultaron en series enteras o mixtas. (Dados en un archivo PDF)
- Las palabras clave son: inversión, balanza (comercial y pagos), PIB, desempleo, inflación, política, fiscal, estructural, monetaria.
- Se conformaron siete apartados:

Indicadores relacionados con	**Número de indicadores (únicos)**
Inversión	5
PIB	21
Balanza comercial	4
Desempleo	5
Inflación	3
Política fiscal	0
Política monetaria	14

3er objetivo: *Graficar de manera eficiente y dar suficientes datos descriptivos de cualquier indicador para un futuro análisis.*

Se automatizó el proceso de búsqueda y de creación de gráficas con tres funciones ***adaptables*** creadas con Python y las librerías *pandas, numpy y matplotlib*.

Buscador() Grafica() GraficaDes()

Datos descriptivos

Cuando un dato no está registrado, la entrada que le corresponde está dada en el formato ***NaN.*** Este formato es especial para indicar *vacío*. Por lo que realizar operaciones fundamentales con este formato no es posible.

De cada serie de datos se extrajeron los datos *registrados* o los *no nulos*, por lo que para determinadas series y sus valores descriptivos pueden estar dados en series mixtas intermitentes. Cabe decir que no son la minoría.

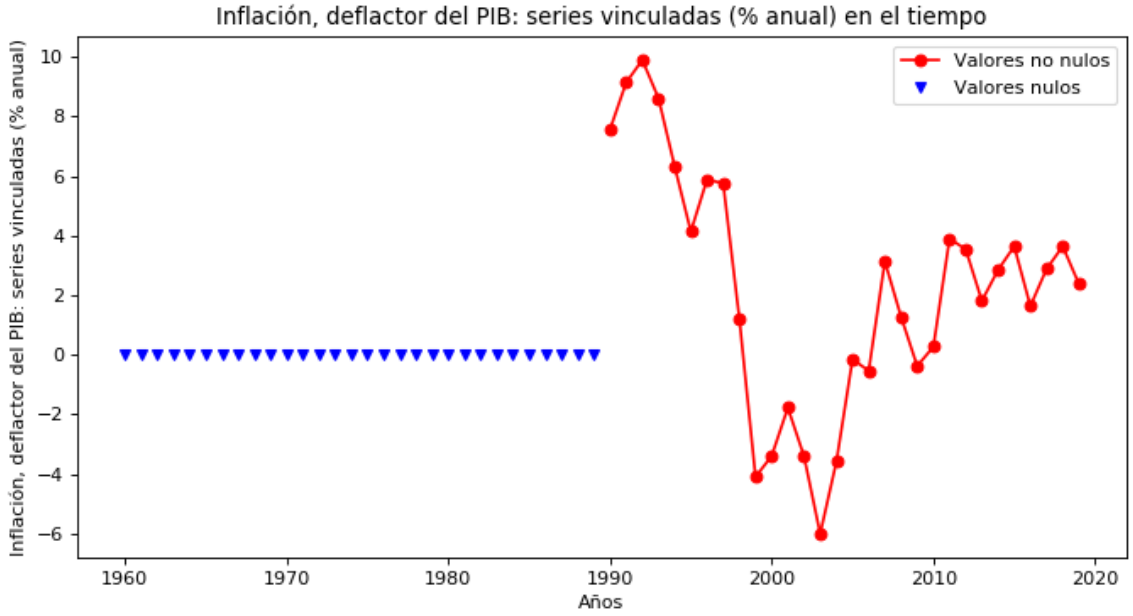
- Datos descriptivos sin datos nulos en las series enteras y mixtas: media, desviación estándar, máximo, mínimo, porcentaje de valores no nulos respecto al total de la serie, número de datos no nulos, serie de tiempo.

Gráficas

Dado que consideramos los registros como variables continuas:

- Representación continua mediante líneas resaltando los puntos de registro con un círculo.
- Representación discreta con triángulos de los valores nulos.

- La representación de los valores nulos está dada en cero solo para fines visuales.
- Cuenta con legenda y etiquetado de los ejes.
- Se indica el tipo de valor graficado en el eje y.
- La dimensión de la gráfica está en la esquina superior izquierda (Depende del valor medido).
- Dentro del entorno de desarrollo son gráficas interactivas.



Ejemplo donde se ilustran todos los elementos propuestos de una gráfica (Hong Kong)

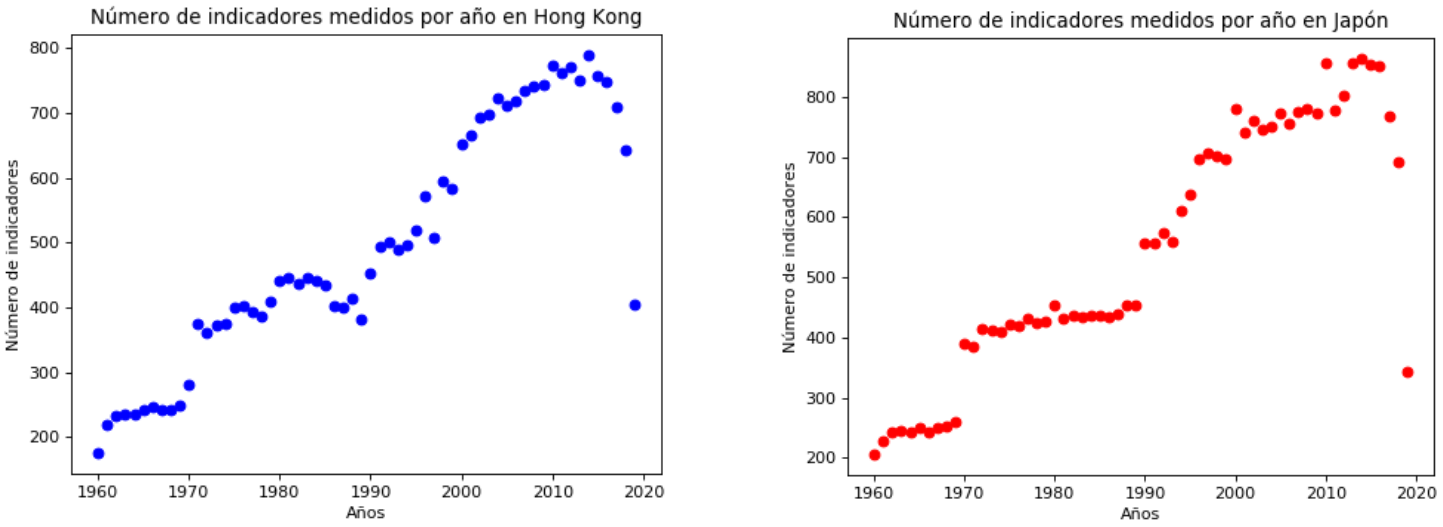
```
Información del Dataset
<class 'pandas.core.frame.DataFrame'>
Index: 60 entries, 1960.0 to 2019.0
Data columns (total 1 columns):
Inflación, deflactor del PIB: series vinculadas (% anual)    30 non-null object
dtypes: object(1)
memory usage: 3.4+ KB
None
-------------------------------------------------
Porcentaje de datos no nulos (%):
Indicator Name
Inflación, deflactor del PIB: series vinculadas (% anual)    50.0
dtype: float64
-------------------------------------------------
Variables descriptivas con datos no Nulos

Desviación estándar con valores no Nulos: 3.9984982359028964
Máximo de la serie: 9.89933756505883
Mínimo de la serie: -6.00770059400513
Media de la serie: 2.20673837422381
-------------------------------------------------
Descripción del Dataset
Indicator Name  Inflación, deflactor del PIB: series vinculadas (% anual)
count                                                       30.000000
unique                                                      30.000000
top                                                         -1.774011
freq                                                         1.000000
-------------------------------------------------
```

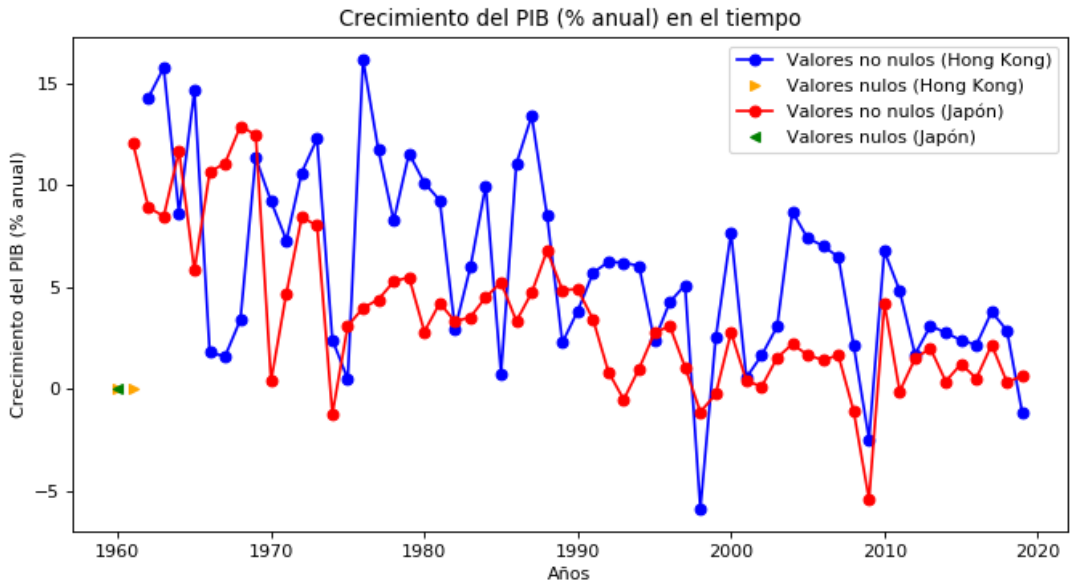
Ejemplo de información desplegada al graficar

4to objetivo: *Graficar y dar datos descriptivos de los indicadores entre Hong Kong y Japón para un futuro análisis.*

- Se buscó que la base de datos de ambos países fuese homogénea.
- Se comprobó que se tiene el mismo número de indicadores e idénticos.
- El proceso para la limpieza del marco de datos de Japón fue el mismo para el de Hong Kong. (ACP)
- La gráfica tiene dos series simultáneas con cuatro legendas.
- La gráfica cumple los elementos descritos para el modelo anterior.
- Los datos descriptivos de ambos países aparecen en el mismo lugar.



Comparación de la medición de indicadores por país por año (Gráficas individuales)

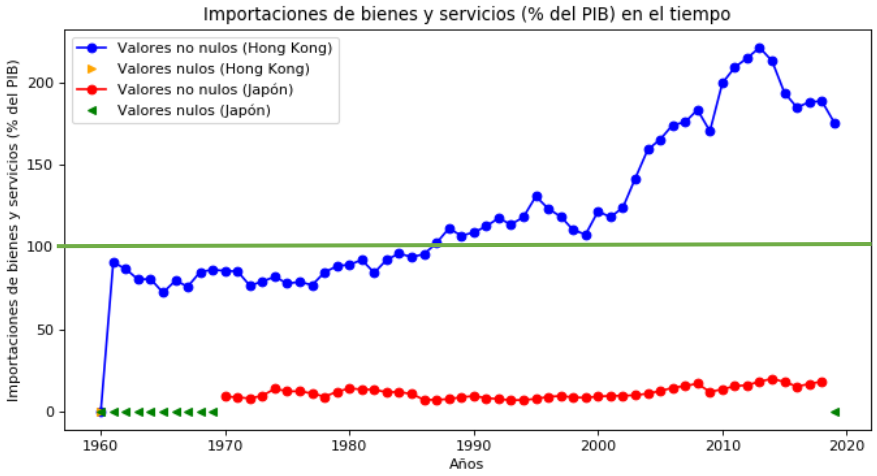


Ejemplo de Gráfica simultánea

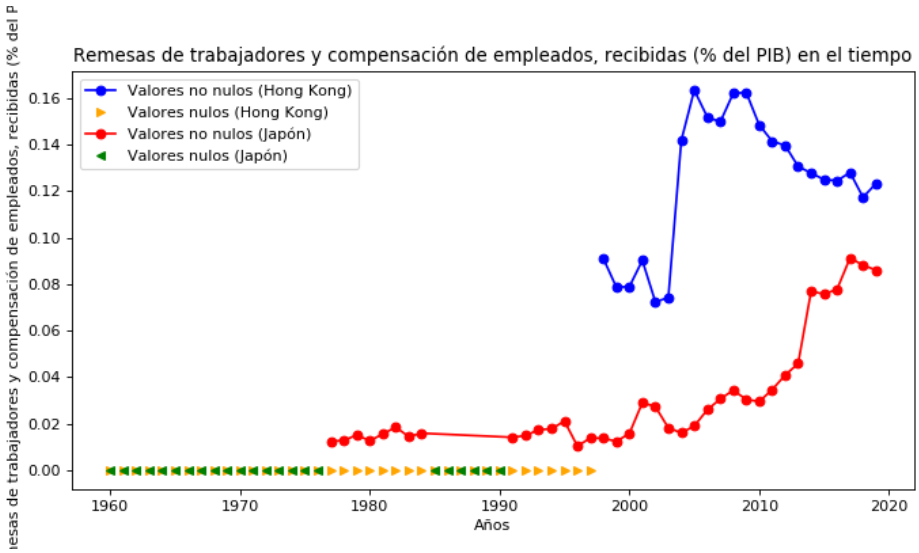
Estado actual *->* ***5to objetivo:*** *Completar las series de datos mixtas (relevantes) menores a 30 registros continuos con datos de otras fuentes, métodos aceptados de sustitución o regresiones para poder tener correlaciones aceptables entre indicadores.*

Retos

1) Comprobar que, aunque los indicadores entre ambos países son idénticos en nombre, el registro tenga las mismas dimensiones o tope porcentual.



En este caso se observa que *importaciones de bienes y servicios (% del PIB),* que debería ser un porcentaje menor que el 100% (pues es sola una parte del PIB), está en esta gráfica por encima del tope en el caso de Hong Kong.



En este caso se observa que *Remesas de trabajadores y compensación de empleados, recibidas (% del PIB)* también está medida en porcentajes y en un rango menor a 100%. Algo matemáticamente más aceptable de inicio.

2) Hay series de datos que, a pesar de ser series mixtas, solo deben de tener un par de registros. De no encontrar datos que tengan una similitud con los anteriores, para evitar lo más que se pueda el sesgo, deberíamos analizar si vale la pena la regresión sobre esa serie. Dependerá de los otros estados del diagnóstico.

3) Debemos intentar llegar a una solución lo antes posible para los datos nulos y tener cuidado de medir el error de cada compensación de estos.